Why log ratios are useful for tracking COVID-19

Rob Hyndman edited by John Minter

Started 2020-04-20, Last modified: 2020-04-20

Contents

Introduction

There have been some great data visualizations produced of COVID-19 case and deaths data, the best known of which is the graph from John Burn-Murdoch in the Financial Times. To my knowledge, it was first used by Matt Cowgill from the Grattan Institute, and has been widely copied. This is a great visualization and has helped introduce log-scale graphics to a wide audience.

Reproducing the Financial Times cumulative confirmed cases graph

To produce something like it, we can use the tidycovid19 package from Joachim Gassen:

```r
library(tidyverse)
library(tsibble)
library(tidycovid19)
```
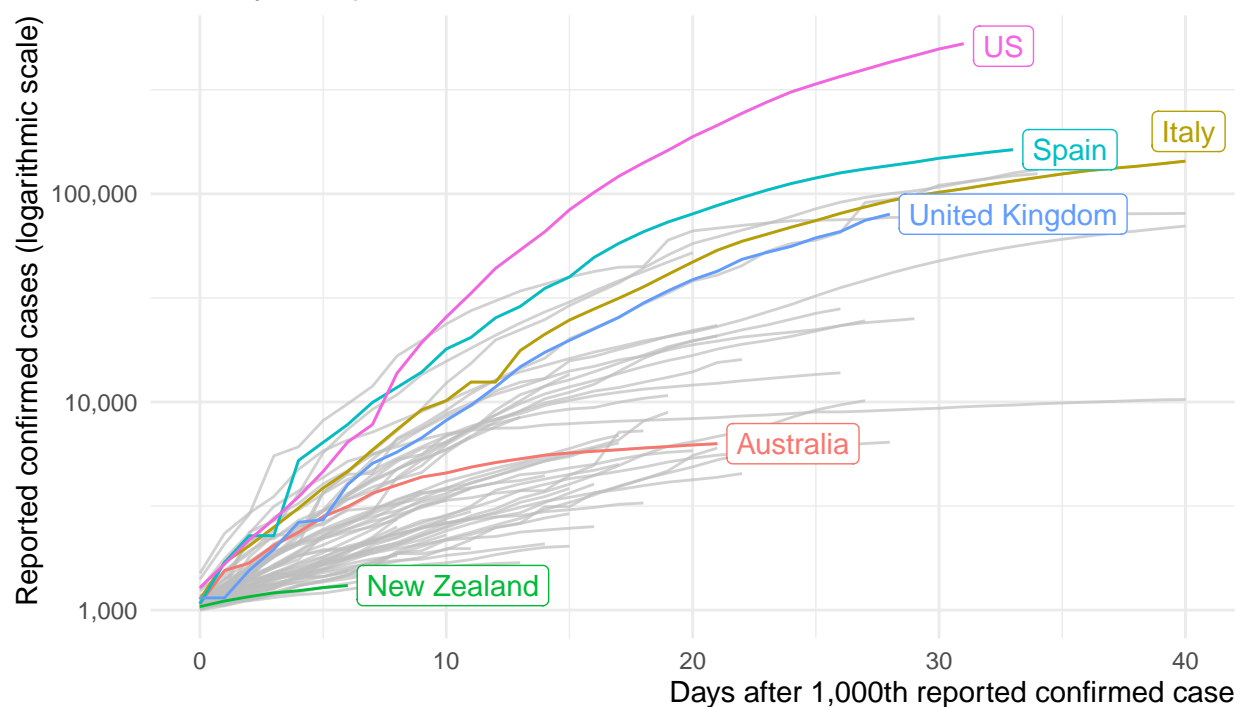
Download latest data

```r
updates <- download_merged_data(cached = TRUE)
```

Plot the country data

```r
# Countries to highlight
countries <- c("AUS", "NZL", "ITA", "ESP", "USA", "GBR")
updates %>%
  plot_covid19_spread(
    highlight = countries,
    type = "confirmed",
    edate_cutoff = 40
  )
```

The First 40 Days: Reported confirmed cases

If you want to produce this from scratch, rather than use the tidycov19 package, see Kieran Healy's post on "Covid 19 Tracking". He also has a nice post on how to produce some of the other plots produced by John Burn-Murdoch.

Why not per capita numbers?

There has been some discussion online as to why per capita numbers aren't shown instead. There are good arguments for using total numbers rather than per capita rates, discussed in this twitter thread. But I wanted to point out something that is often overlooked.

Per capita numbers would only shift the curves vertically, they would not change the shape or slopes of any curves. To see this, remember that

$$log(Y_t/P_t) = log(Y_t) - log(P_t).$$

Here Y_t is the number of cases on day t and P_t is the population on day t. The population hardly changes over a period of a few weeks or months, so taking per capita numbers is simply subtracting a constant from the log cases data.

So if we just concentrate on the slope of those lines, not on their vertical position, we are effectively adjusting for population.

Differences in testing regimes

A similar point applies to the differences in testing regimes between countries. If Country 1 has a testing regime that only identifies p % of cases, and Country 2 has a testing regime that identifies q % of cases, then these differences will show up as a vertical difference between the curves - they won't affect the slopes.

2

Of course, countries have been changing their testing regimes over time, so changes in the slope of the curve could be a result of a change in the rules of who gets tested, or they could be the result of an reduction in the total number of cases. There is no way to tell from the data which of these is likely to be true.

Focus on log ratios

All of this shows that we should be looking at the slope of these lines if we are wanting to explore how countries are tracking relative to each other. That is equivalent to looking at log ratios, since the slope of the log curve is the same as the log of the ratios of successive values.

First let's compute the log ratios and plot them for each country with a smooth curve overlaid.

```
updates %>%
  mutate(cases_logratio = difference(log(confirmed))) %>%
  filter(
    iso3c %in% countries,
    date >= as.Date("2020-03-01")
  ) %>%
  ggplot(aes(x = date, y = cases_logratio, col = country)) +
    geom_point() +
    geom_smooth(method = "loess") +
    facet_wrap(. ~ country, ncol = 3) +
    xlab("Date") +
    ggthemes::scale_color_colorblind()
```
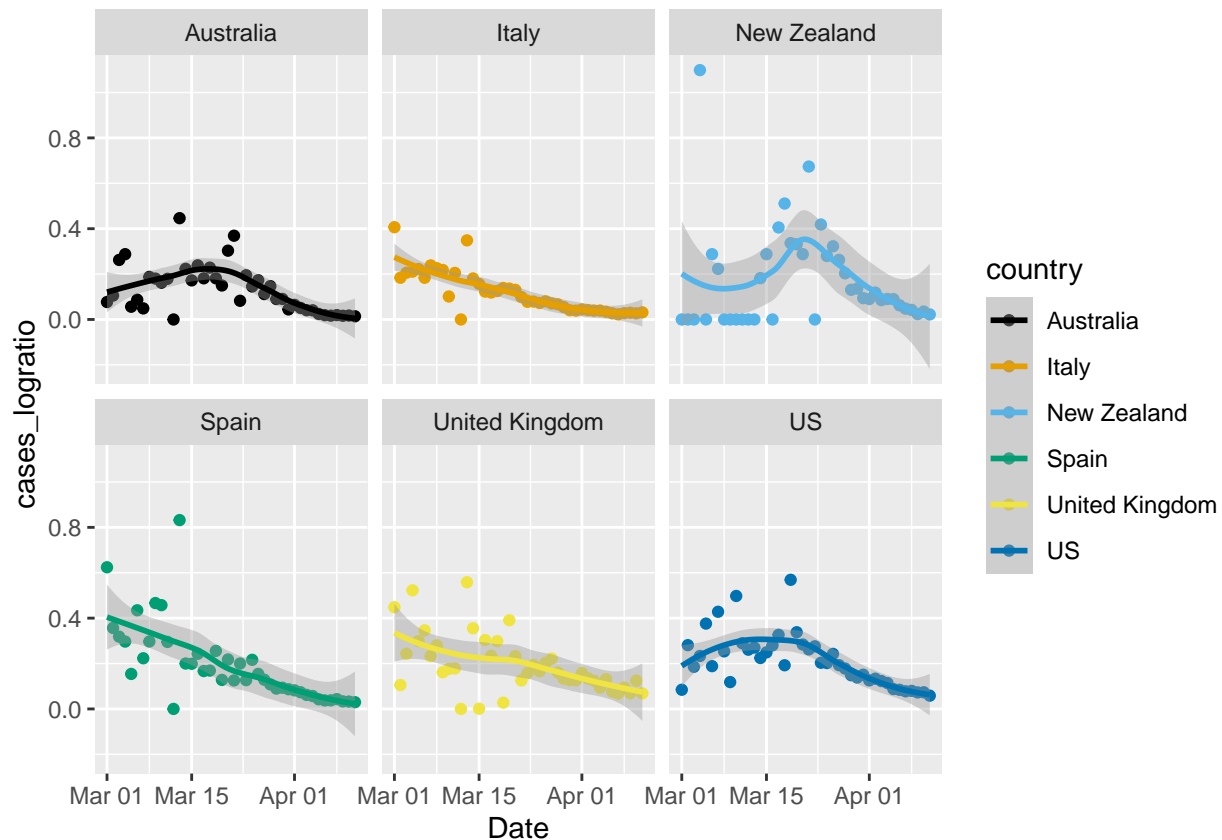

There are some obvious reporting issues (e.g., Spain had no new cases on 12 March, but roughly twice as many cases as we might expect on the 13 March), but the smooth (`loess`) curve corrects for these discrepancies. We can plot the smoothed log ratios to see how each country is tracking in terms of the slope of the curve.

```
updates %>%
  mutate(
    cases_logratio = difference(log(confirmed))
  ) %>%
  filter(iso3c %in% countries) %>%
  filter(date >= as.Date("2020-03-01")) %>%
  ggplot(aes(x = date, y = cases_logratio, col = country)) +
  geom_hline(yintercept = log(2)/c(2:7,14,21), col='grey') +
  geom_smooth(method = "loess", se = FALSE) +
  scale_y_continuous(
    "Daily increase in cumulative cases",
    breaks = log(1+seq(0,60,by=10)/100),
    labels = paste0(seq(0,60,by=10),"%"),
    minor_breaks=NULL,
    sec.axis = sec_axis(~ log(2)/(.),
      breaks = c(2:7,14,21),
      name = "Doubling time (days)")
  ) +
  ggthemes::scale_color_colorblind()
```
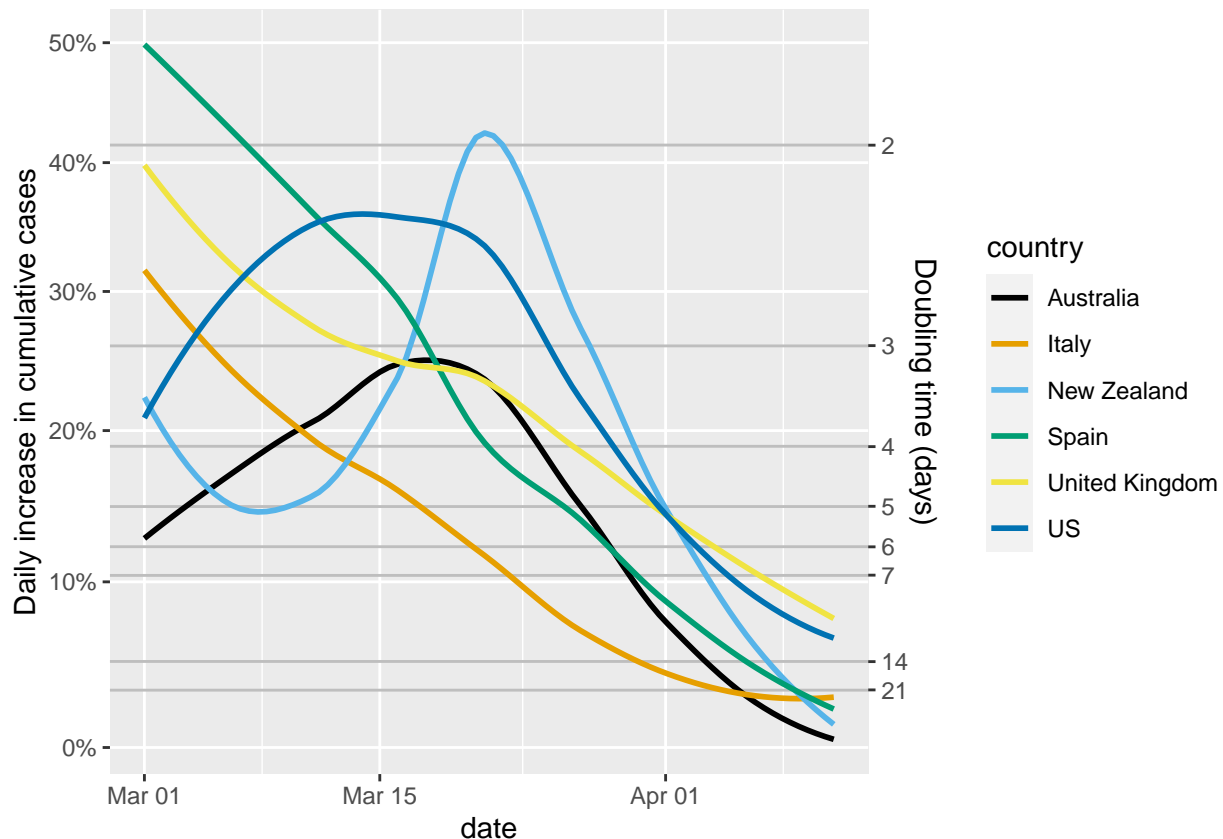

I've added axes that are a little more interpretable than log-ratios. On the left axis we have daily increase in cumulative cases. If r is the log ratio, then $100(e^r - 1)$ is the daily increase as a percentage. On the right axis we have the number of days it takes for the cumulative cases to double. This is easily computed as $d = log(2)/r$.

While the graph doesn't tell us anything about how the disaster is overwhelming the health systems of each country (the cumulative case graph at the top of this post is best for that), it does tell us how well each

country is managing to slow the rate of infection and reduce the slope of those curves. It is also much easier to compare the slopes this way, because we just have to look to see which country is higher.

In a country like Italy where the hospitals are already overwhelmed with cases, any continuing daily increase is much more serious than in Australia and New Zealand where the numbers of cases are still relatively low compared to much of Europe and the US. For Australia and New Zealand, if we can keep the daily increases where they currently are, we can help spread the load on the health system and save lives.